



Date published: 16 October, 2024

Date last updated: 16 October, 2024

Planning and implementing real-world artificial intelligence (AI) evaluations: lessons from the AI in Health and Care Award

[Publication \(/publication\)](#)

Content

- [Introduction](#)
- [The approach to developing this document and future work](#)
- [How the evaluations worked](#)
- [Lessons for national teams designing programmes to support real-world evaluation of new technologies](#)
- [Detailed review of lessons learned at each stage of the evaluation process](#)
- [Acknowledgements](#)
- [Appendix A - case studies](#)
- [Appendix B: evaluation questions and associated methods](#)
- [Appendix C: national guidance and resources](#)
- [Appendix D: theoretical frameworks for the evaluation of AI technologies in health and care](#)
- [Appendix E: evaluations covered by this review](#)

Introduction

The [Artificial Intelligence \(AI\) in Health and Care Award](https://www.england.nhs.uk/aac/what-we-do/how-can-the-aac-help-me/ai-award/) (<https://www.england.nhs.uk/aac/what-we-do/how-can-the-aac-help-me/ai-award/>) ('AI Award') ran from 2020 to 2024 and was part of the [NHS AI Lab](https://transform.england.nhs.uk/ai-lab/) (<https://transform.england.nhs.uk/ai-lab/>), a Department of Health and Social Care initiative included in the Government Major Projects Portfolio. It allocated more

than £100 million to support the design, development and deployment of promising AI technologies.

The AI Award was structured into 4 'phases' based on how ready products were for real-world implementation and the evidence available to support wider adoption:

- Phase 1: AI design, training and feasibility testing
- Phase 2: further AI development and clinical validation
- Phase 3: first prospective real-world deployment of the AI
- Phase 4: multi-site deployment and real-world evaluation

The 13 Phase 4 technologies from the first and second rounds were independently evaluated with the aim of informing assessment by the National Institute for Health and Care Excellence (NICE) and the UK National Screening Committee, as well as local and national commissioning decisions.

This document looks at what was learned from these Phase 4 evaluations, with a specific focus on the evaluation process itself.

It provides lessons on the practical 'how to' of designing and implementing real-world evaluations of AI and is intended to sit alongside [theoretical frameworks for evaluation](https://www.england.nhs.uk/long-read/planning-and-implementing-real-world-ai-evaluations-lessons-from-the-ai-in-health-and-care-award/#appendix-d-theoretical-frameworks-for-the-evaluation-of-ai-technologies-in-health-and-care) (<https://www.england.nhs.uk/long-read/planning-and-implementing-real-world-ai-evaluations-lessons-from-the-ai-in-health-and-care-award/#appendix-d-theoretical-frameworks-for-the-evaluation-of-ai-technologies-in-health-and-care>). It will be useful to teams implementing and evaluating AI in health and care, as well as to national teams working to support innovation.

The approach to developing this document and future work

Between October 2022 and March 2023, the following steps were taken by an independent partner (Unity Insights) in preparation for this publication:

- a high-level literature review on AI evaluations to provide context for the document review, interviews, and focus group discussions
- a review of documents from the projects shown in [Appendix E](https://www.england.nhs.uk/long-read/planning-and-implementing-real-world-ai-evaluations-lessons-from-the-ai-in-health-and-care-award/#appendix-e-evaluations-covered-by-this-review) (<https://www.england.nhs.uk/long-read/planning-and-implementing-real-world-ai-evaluations-lessons-from-the-ai-in-health-and-care-award/#appendix-e-evaluations-covered-by-this-review>). The documents included 13 evaluation plans, the AI Award Evaluation Advisory Group papers and minutes where proposed changes to evaluation plans were discussed and 4 interim results reports.
- interviews and a focus group with evaluation teams covering topics such as evaluation design choices, barriers and enablers and their learnings to-date

- insight on the challenges of and approaches to real-world AI evaluation was generated according to the 4 stages of the evaluation lifecycle (scoping, designing and planning, conducting the evaluation, and disseminating the findings).

Due to the timing of this review, only 4 of the 13 projects had interim results. The projects with interim results were all Round 1 projects and the Round 2 projects were either still in the 'Design and Plan' stage or had only just moved into the 'Conduct' stage. Most of the insights therefore relate to the 'Scoping' and 'Designing and Planning' stages.

The [NHS AI Lab](https://transform.england.nhs.uk/ai-lab/) (<https://transform.england.nhs.uk/ai-lab/>) plans to publish an evaluation of all NHS AI Lab programmes. It will gather insights from all the AI Award project reports to spotlight benefits and identify further lessons.

How the evaluations worked

A [NICE MedTech Early Technical Assessment \(META\)](https://meta.nice.org.uk/) (<https://meta.nice.org.uk/>) was commissioned for each Phase 4 technology where NICE was the key audience for the evaluation.

These META assessments were then used to guide the design of each independent evaluation, which aimed to fill in evidence gaps highlighted in the META.

The evaluations examined each AI technology across 8 evaluation 'domains' as they were used in real-world healthcare settings:

1. **Safety:** what were the key risks associated with the technology and what risk assurance or management was in place?
2. **Accuracy:** how accurate was the AI technology in a real-world environment?
3. **Effectiveness:** what were the clinical, social, experiential or operational effectiveness outcomes?
4. **Value:** What was the health economic and budgetary impact of the AI technology?
5. **Fit with sites:** was the technology addressing requirements and population needs at sites where it was deployed? Was the technology acceptable to clinicians, patients and the public?
6. **Implementation:** what were the reasons for implementing the AI at sites where it was deployed – and what were the barriers?
7. **Feasibility of scaling up:** what was the scale-up strategy? How would differences in IT infrastructure and clinical practice between sites be accommodated?

8. **Sustainability:** how would sustainable use of the technology be ensured (for example, pricing, incorporation of customer feedback, ongoing model training)?

The evaluation teams were also asked to look at how patient characteristics influenced the results in each domain and whether some groups of patients might miss out or be negatively affected.

Lessons for national teams designing programmes to support real-world evaluation of new technologies

1. National oversight of designs will help ensure stakeholder expectations are met

The quality of the evaluation designs was assured by an Evaluation Advisory Group (<https://www.england.nhs.uk/aac/what-we-do/how-can-the-aac-help-me/ai-award/evaluation-for-artificial-intelligence/>), which included representatives from NICE, the UK National Screening Committee, academics specialising in AI regulation, adoption and evaluation, and experts in patient and public involvement and engagement (PPIE).

This input – as well as input from national programme teams in specific clinical areas – helped ensure the evaluations were fit for purpose.

2. AI deployment and evaluation plans must be co-produced by technology suppliers, independent evaluators, and adopting sites (including clinical and patient users)

The independent evaluators helped ensure accurate assessment of the accuracy of algorithms and identified opportunities for improvement. However, issues were encountered in 2 areas:

- The independent evaluations were commissioned after projects had been selected as part of the awards process. This meant the deployment plans of the technology companies were already decided and this sometimes limited the evaluations (affecting, for example, the selection of sites, the feasibility of randomisation, and the time available for data collection).
- The AI Award projects were led by the technology suppliers rather than adopting sites so sometimes encountered issues with a lack of capacity or motivation from sites to participate in the AI deployments or evaluations. Some projects didn't sufficiently consider how the technology would integrate into clinical pathways and what its down-stream impact would be.

4. At least 2 years should be allowed for evaluations of multi-site technology deployments

Several months each were required for:

- designing and planning the evaluation, including gaining required research approvals and contracting with sites
- understanding the baseline situation (and outcomes) and understanding variation between sites
- bedding the technology in before starting to measure impact

5. Future national programmes should encourage quasi-experimental, mixed-method evaluation designs

Only 2 of the 13 evaluations were designed as randomised controlled trials. The rest were quasi-experimental (<https://www.gov.uk/guidance/quasi-experimental-study-comparative-studies>) pre-post implementation studies. Quasi-experimental approaches were more suited to AI implementation, allowing for rapid updates to the technology platforms.

The impact of the AI technologies was also highly dependent on the context in which they were deployed. Mixed methods (<https://www.gov.uk/guidance/mixed-methods-study>) evaluation designs were effective in helping teams understand these effects.

This is in line with the Health Foundation's advice on the testing and evaluation of AI (<https://www.health.org.uk/publications/long-reads/priorities-for-an-ai-in-health-care-strategy>).

5. Future national programmes should consider a more explicit focus on assessing the impact of AI technologies on health inequalities (for example, by including a stand-alone evaluation domain)

The potential for AI algorithms used in healthcare to produce biased outputs and exacerbate health inequalities is an area of concern. Some evaluators planned to include sensitivity analyses as part of their evaluations to help them understand if the overall outcomes observed varied according to patient characteristics. However, this was not done consistently.

6. Rapid changes in the sector means teams lean heavily on national guidance and resources – and need this information to be regularly updated

The guidance and advice came from national organisations including the Medicines and Healthcare products Regulatory Agency (MHRA), Health Research Agency (HRA), National Institute for Health and Care Excellence (NICE), NHS England, and the government.

See [Appendix C \(https://www.england.nhs.uk/long-read/planning-and-implementing-real-world-ai-evaluations-lessons-from-the-ai-in-health-and-care-award/#appendix-c-national-guidance-and-resources\)](https://www.england.nhs.uk/long-read/planning-and-implementing-real-world-ai-evaluations-lessons-from-the-ai-in-health-and-care-award/#appendix-c-national-guidance-and-resources) for key resources.

Detailed review of lessons learned at each stage of the evaluation process

The scoping stage

The evaluation teams were all required to scope their evaluations as part of their applications to take part in the awards. Successful teams were then given 2 months alongside the technology companies whose technologies they were evaluating to refine this work. The scope was then reviewed by the [Evaluation Advisory Group \(https://www.england.nhs.uk/aac/what-we-do/how-can-the-aac-help-me/ai-award/evaluation-for-artificial-intelligence/\)](https://www.england.nhs.uk/aac/what-we-do/how-can-the-aac-help-me/ai-award/evaluation-for-artificial-intelligence/).

3 key lessons were learned relating to the scoping stage:

1. Establish the regulatory status of the AI technology, its intended purpose, and value claims at the outset – and identify any plans that may change these characteristics

Evaluation teams said it was important to build a close relationship with the technology company during the scoping phase and to deeply understand their technology, its purpose and value claims, and its regulatory status. Understanding the company's strategic plans for the technology and whether this might affect the evaluation was also crucial.

For example, one technology company planned to apply for a higher medical device certification level during the AI Award (from Class I to Class IIA UKCA mark). A good relationship between the evaluation team and technology company supported early and frank discussions about the technology's regulatory status and identification of the most appropriate evaluation design and evidence requirements. This avoided the need to change the evaluation 'in flight'.

There were 2 specific issues preventing clarity at this stage for some projects:

- Differing opinions over whether AI algorithms whose intended purpose was

to improve operational efficiency of clinical administration tasks were medical devices. This was because there was uncertainty over whether improving operational efficiency in these cases counted as a medical purpose, as outlined by the [Medicines and Healthcare Products Regulatory Agency \(MHRA\)](https://www.gov.uk/government/publications/medical-devices-software-applications-apps) (<https://www.gov.uk/government/publications/medical-devices-software-applications-apps>). Where uncertainty exists technology companies and evaluators are encouraged to engage regulatory expertise.

- Differing opinions in some projects between the evaluation team, the technology company, and NICE on the level of risk that AI algorithms with 'operational efficiency' use cases posed to patient care if they went wrong. This was relevant to which tier of the [NICE Evidence Standards Framework](https://www.nice.org.uk/corporate/ecd7/chapter/section-b-classification-of-digital-health-technologies) (<https://www.nice.org.uk/corporate/ecd7/chapter/section-b-classification-of-digital-health-technologies>) the technologies would fall into.

2. Engage with sites early and give sufficient time to understand the variation between sites' clinical practices and IT systems and their capacity to participate in the evaluation

It was important to understand the capacity and appetite of local clinical and research and development teams to implement the AI technologies and to collect, clean and share data around the use and outcomes associated with the new technology. Any expected changes to this status also needed to be grasped early.

For example, one trust was planning to upgrade their electronic patient record system during the timeframe of the award. This reduced the site's capacity to participate in the evaluation and could have affected the results of the pre-post evaluation.

Changing clinical practice was also significant. One evaluation had to be rescoped part way through the project because a drop in the use of a certain kind of test for diagnosing cardiovascular diseases resulted in fewer cases than expected.

3. Contact the Health Research Authority if it is unclear whether an evaluation constitutes 'research' or a 'service evaluation'

Evaluators that classified their evaluation as 'service evaluation' rather than research sometimes encountered problems. Research requires approval by the Health Research Authority (HRA), whereas service evaluations, which are conducted as part of routine care, don't.

Some projects, which were focused on medical devices that had marketing

approval (CE or UKCA marking) or technologies that they did not consider to be medical devices, classified their work as service evaluation but then found that national dataset managers and site research and development offices would not allow data access until proof of HRA approval had been provided.

Written confirmation from the HRA that the evaluation was a service evaluation was useful in overcoming these hurdles. The HRA has [an online tool to help teams find out whether studies constitute research \(https://hra-decisiontools.org.uk/research/\)](https://hra-decisiontools.org.uk/research/), but, where there is uncertainty, advice from the HRA can be sought through the [AI and Digital Regulations Service \(https://www.digitalregulations.innovation.nhs.uk/support-for-developers/\)](https://www.digitalregulations.innovation.nhs.uk/support-for-developers/) or by contacting [queries@hra.nhs.uk \(mailto:queries@hra.nhs.uk\)](mailto:queries@hra.nhs.uk).

Designing and planning the evaluation

The evaluation teams had 2 months after the Evaluation Advisory Group approved their evaluation scopes to work with the technology companies, site stakeholders, and patient and public involvement and engagement (PPIE) representatives to further refine their 'theory of change', update evaluation questions and confirm the more detailed methods and analytical approaches that would be used to answer the evaluation questions.

In addition, evaluation teams were expected to develop their detailed project plan, including a detailed timeline, milestones, deliverables, governance, the division of roles and responsibilities between project partners, risk management plan and stakeholder management plan.

After the evaluation plans were approved, evaluation teams went on to develop and submit research protocols for approval by the HRA (where relevant). They also engaged with sites to further understand capacity and motivation to take part in the studies and enter into collaboration and data-sharing agreements with sites, where required. Collaboration agreements were entered into between the evaluation team and the technology company, dividing roles and responsibilities and clarifying data-sharing arrangements.

8 key lessons were learned relating to this stage:

1. Mixed qualitative and quantitative methods are crucial to understanding impact

All of the evaluation teams used mixed qualitative and quantitative methodologies to examine technologies across the 8 evaluation domains. [Appendix B \(https://www.england.nhs.uk/long-read/planning-and-implementing-real-world-ai-](https://www.england.nhs.uk/long-read/planning-and-implementing-real-world-ai-)

[evaluations-lessons-from-the-ai-in-health-and-care-award/#appendix-b-evaluation-questions-and-associated-methods](#)) summarises the questions evaluations sought to answer and the methods used.

This [mixed methods](https://www.gov.uk/guidance/mixed-methods-study) (<https://www.gov.uk/guidance/mixed-methods-study>) approach helped evaluation teams to understand the variety of factors affecting clinical outcomes. The implementation of the AI sometimes drove or accompanied changes to clinical practice, so the qualitative methods were important to understanding the technology's contribution to impact. For example, one AI technology analysed CT images and supported clinical decisions about treatments for stroke patients. However, it was embedded in a platform that also enabled remote access to the CT images by radiologists. The qualitative component of the evaluation allowed the evaluators to assess the impact of the AI-enabled component and separate it from the impact of image sharing.

The use of qualitative as well as quantitative approaches allowed evaluators to understand variation between sites that might affect outcomes, such as:

- differences in clinical pathways and care practices before the evaluation
- differences in local IT systems and how they are used
- variations in the calibration of AI technologies
- differences in access to the AI (or similar technology) and to training prior to the evaluation
- variations in the use of the AI outputs by clinicians or administrative teams

The qualitative information also provided valuable insights into the accuracy of the technologies. For example, one evaluator noted that radiologists found that nodules were difficult to pick up by the technology if they were next to a blood vessel.

2. Pragmatic and flexible evaluation designs work best

Only 2 of the 13 evaluations were planned as randomised controlled trials (both were [cluster-randomised trials](https://www.bmj.com/content/358/bmj.j3064) (<https://www.bmj.com/content/358/bmj.j3064>) at site level). The rest were [quasi-experimental](https://www.gov.uk/guidance/quasi-experimental-study-comparative-studies) (<https://www.gov.uk/guidance/quasi-experimental-study-comparative-studies>) pre-post implementation evaluations.

Some technologies had already been implemented before the awards, making randomisation impossible, and the quasi-experimental approaches were also more adaptable if sites no longer had the capacity to take part in studies. Some projects didn't want to delay implementation of the technologies for randomisation and assignment (in the case of [stepped wedge cluster-randomised trials](https://www.bmj.com/content/350/bmj.h391) (<https://www.bmj.com/content/350/bmj.h391>)) or for trial completion (in the case of other

randomised controlled trials).

The quasi-experimental approach also had the advantage of flexibility in a very rapidly changing AI environment. This allowed for updates to technology platforms (for example, improvements in the accuracy of an AI algorithm due to re-training, improvements in the user interface, or to its cyber security features).

To increase confidence in the conclusions from pre-post intervention comparisons, some of the evaluations also included outcome data from a matched control site where the technology had not been implemented. This allowed comparison of outcomes and provided more information on variables such as seasonality and changes in clinical practice.

3. Access to independent AI expertise and clinical leadership is valuable

Effective evaluation typically required multidisciplinary teams with expertise in quantitative methods, qualitative methods, health economics, information governance, and project management. 2 additional areas of expertise were found to be important:

- independent AI or machine learning expertise
- relevant clinical understanding

As part of the “accuracy” domain, a number of the evaluations required independent assessment of algorithm design and training methodologies by a machine learning expert or independent assessment of the performance of the algorithm compared to clinical experts. Some found accuracy results that differed from those reported by the technology company. Opportunities for improvement in algorithm accuracy (using alternative algorithm training and validation techniques) were also identified.

The inclusion of clinical leadership in project teams was also vital. The AI technologies had broad use cases and so could be included in clinical pathways in different ways (but still within the regulatory-approved purpose). For example, an AI technology used to detect potential cancer from images could be used to replace a human reader in the screening workflow or could be used as a safety net to flag potential cancer cases after the human readers’ review. Clinical leadership was essential to designing the AI-enabled workflow, determining what priority outcomes should be measured, and considering unintended consequences of the AI implementation.

4. The evaluation audience should be consulted when establishing the

approach to health economic modelling

The evaluators used a variety of different health economic modelling approaches to help them understand the value of the AI technologies. 5 looked at budget impact (<https://www.gov.uk/guidance/budget-impact-analysis-health-economic-studies>), 4 looked at cost-effectiveness (<https://www.gov.uk/guidance/cost-effectiveness-analysis-health-economic-studies>), 3 used a cost benefit analysis (<https://www.gov.uk/guidance/cost-benefit-analysis-health-economic-studies>), 3 used a cost utility analysis (<https://www.gov.uk/guidance/cost-utility-analysis-health-economic-studies>) and 1 employed a cost-consequence (<https://www.gov.uk/guidance/cost-consequence-analysis-health-economic-studies>) approach.

Where the main audience for the evaluation was NICE, evaluators used the NICE Evidence Standards Framework for Digital Health Technologies (<https://www.nice.org.uk/corporate/ecd7/chapter/how-to-meet-the-standards#delivering-value>) for guidance on the most appropriate approach.

Where NICE was not the main audience (for example, for technologies being incorporated into the breast- or eye-screening pathways or in trust operational efficiency processes) a steer from the national screening programme teams or trust procurement teams was sought.

5. Dialogue about and written agreement on data collection and sharing arrangements to enable the evaluations between the evaluation partner, the technology company, and the sites are critical

Differing opinions on information governance roles and responsibilities between the evaluation partner, the technology company and sites were key sources of delay in a number of the evaluations.

Areas of disagreement included:

- whether the technology company would be able to share technology use and outcome data with the evaluator or whether the evaluator needed to source the data directly from the sites
- how data anonymisation would be handled (for example, whether the sites would anonymise the data and then pass it to the evaluation team, whether the company would anonymise the data then pass it to evaluators, or whether the evaluation partner would receive pseudonymised data directly from the sites)

The key factors in avoiding issues were early dialogue, early written agreement

on roles and responsibilities, and ensuring sufficient information governance expertise and capacity within evaluation teams.

Early conversations with sites were also critical to understanding the feasibility of data collection (and to updating evaluation methods and metrics accordingly).

Teams needed to understand:

- what caseloads were anticipated at sites to establish how much time or what number of sites were needed to collect the required data
- whether outcome data could be sourced from routinely collected data or if primary data collection was needed
- what limitations might be encountered if outcome data needed to be sourced from routinely-collected data (for example, coding errors or missing data) and what curation might be required (for example, analysis of free-text fields)
- what resource would be required at sites to collect, curate and transfer the data to the evaluators

There were often differences between the caseloads and the data collection and data sharing arrangements at sites (for example, some sites required the same data-sharing agreements to be signed every year). Where primary data collection was required, evaluation teams also sometimes found that the technology suppliers could support with data collection (for example, by including a pop-up to ask clinicians how much they agreed with the AI algorithm's analysis of CT images for the detection of lung cancer nodules).

However, some evaluators found that engagement with sites about the design and planning of the evaluation had to be iterative. A high-level plan was initially brought to sites to gauge appetite, capacity, and the feasibility of data collection. The evaluators then had to return to sites to validate their updated plans, once thinking had evolved. It was important to get written agreement about roles and responsibilities of the evaluation team and the technology company early, but full agreement between evaluation teams and individual sites sometimes waited until the final stages of the design and planning stage when the detail of the evaluation was clear.

6. Involvement of patient 'super-users' or patient organisations can be helpful when designing evaluations of clinician-facing AI technologies

Evaluators sometimes had difficulty with patient engagement when dealing with diagnostic technologies that did not directly involve patients. For example, a patient helped by an AI technology supporting the detection of cancerous lesions based on image analysis might not be aware of the technology or have any

interaction with it.

One evaluation team steered their PPIE approach toward ‘super users’ and patient groups with organisational relationships with deployment sites and who consistently work with the NHS. These groups had more experience and understanding of digital projects.

7. Several months should be allowed for baseline analysis and to allow technologies to bed in before measuring impact

When an evaluation was comparing pre and post deployment outcomes, it was important to set aside several months to collect data and understand the pre-deployment clinical pathways. This was especially important if there was likely to be significant variation in practice between sites or IT systems (for example, between electronic patient record systems; see [Appendix A \(https://www.england.nhs.uk/long-read/planning-and-implementing-real-world-ai-evaluations-lessons-from-the-ai-in-health-and-care-award/#appendix-a-case-studies\)](https://www.england.nhs.uk/long-read/planning-and-implementing-real-world-ai-evaluations-lessons-from-the-ai-in-health-and-care-award/#appendix-a-case-studies), case study 1).

Allowing several months for new technologies to bed in before measurement helped evaluators avoid underestimating their impact. This period allowed time for the integration and calibration of the AI technology at sites (for example, modifying the sensitivity and specificity thresholds based on local populations or clinical practices) and for users to be trained and gain experience.

This bedding-in period can be achieved using parallel running, with old processes continuing as the new technology starts to operate. Clinicians might not be using the technology to guide real decisions, but its recommendations are recorded and compared to routine clinical decisions.

8. Evaluations should include thorough analysis of outcomes in the light of patient characteristics – and of clinicians’ experience

AI algorithms used in healthcare have the potential to produce biased outputs and can exacerbate health inequalities. Missing or inaccurate data in training or validation datasets can lead to under-diagnosis or under-estimation of the healthcare needs in disadvantaged groups. Some evaluators planned to include sensitivity analyses to detect variation in outcomes due to patient characteristics such as race or age. Impressions of the accuracy and fairness of the technologies can also be assessed through qualitative methods.

Another key area of texture was clinicians’ experience. One interim evaluation results report indicated that the AI algorithm was likely to increase confidence in

decision-making by more junior radiologists, but that it was unlikely to provide benefit to senior radiologists who were generally more confident about their interpretation of CT scans. Variations in clinician's opinions of and use of technologies will often be best explored qualitatively rather than quantitatively because of the small numbers of people likely to be involved.

Conducting the evaluation

Evaluation teams established their governance structures, including setting up steering groups and stakeholder engagement mechanisms such as a patient and public involvement and engagement forum (if these had not already been set up during the design and planning stage).

They then started collecting quantitative and qualitative data and analysing it according to their evaluation plan. Frequent data quantity and quality checks helped ensure the required quantitative sample sizes were reached. Economic model building was part of the 'value' workstream of all projects.

Progress with delivering the evaluations was reported to the award's delivery team and some requests were received to amend the scope or evaluation plans of projects due to:

- sites deciding not to deploy the technology or participate in evaluation
- technology use being lower than expected (see [Appendix A \(https://www.england.nhs.uk/long-read/planning-and-implementing-real-world-ai-evaluations-lessons-from-the-ai-in-health-and-care-award/#appendix-a-case-studies\)](https://www.england.nhs.uk/long-read/planning-and-implementing-real-world-ai-evaluations-lessons-from-the-ai-in-health-and-care-award/#appendix-a-case-studies), case study 2)
- insufficient data for analysis because of data quality issues at sites.

2 key lessons were learned relating to this stage:

1. Regular checks on the use of technologies and the quality of data being collected are important – and allow evaluation designs to be flexed

Some technology companies and evaluators found that AI outputs were not being used to the extent or in the way expected (see [Appendix A \(https://www.england.nhs.uk/long-read/planning-and-implementing-real-world-ai-evaluations-lessons-from-the-ai-in-health-and-care-award/#appendix-a-case-studies\)](https://www.england.nhs.uk/long-read/planning-and-implementing-real-world-ai-evaluations-lessons-from-the-ai-in-health-and-care-award/#appendix-a-case-studies), case study 2). This threatened recruitment targets.

Some evaluators also found that the quantity or quality of data they were receiving wasn't sufficient for meaningful analysis. Frequent (at least quarterly)

data quantity and quality checks meant evaluation designs could be changed in response to these issues. Solutions included:

- recruiting additional sites
- extending the data collection period
- switching to a retrospective study (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6176693/>) design rather than a prospective design (see Appendix A (<https://www.england.nhs.uk/long-read/planning-and-implementing-real-world-ai-evaluations-lessons-from-the-ai-in-health-and-care-award/#appendix-a-case-studies>), case study 2)

Note that switching to a retrospective design could mean that an evaluation no longer meets the requirements of the audience (such as NICE), so could result in additional evaluation work being required in the future.

2. Build up a picture of the acceptability of incorporating AI into a care pathway by talking to patients

Some of the evaluations planned to interview patients who had been diagnosed with the support of the AI technologies (for example, those who had had an issue detected by AI analysis of their CT scans). These interviews would help evaluators to understand patient experiences of their diagnosis — and their attitudes to an AI technology being involved. Due to the delays and challenges with the deployment of the technologies, some of the evaluators instead turned to hypothetical questions, asking a representative sample of patients how they would feel if AI was involved in supporting their diagnosis.

Disseminating findings

Dissemination and communications plans were produced during the ‘design and plan’ and ‘conduct’ stages. These often involved additional deliverables beyond formal reporting, including infographics, webinars, conference abstracts, and submissions to academic journals.

Evaluations lasting longer than 1 year were required to report interim results. At the time of the work done for this review, 4 of the round 1 evaluations had done this, reporting findings about:

- the baseline clinical pathways into which the AI would integrate and variation in practice between sites
- barriers to and enablers of technology deployment (process evaluation results)
- trends in early impact

- literature review results and details of their planned health economic modelling approaches.

The reports were reviewed by the [Evaluation Advisory Group \(https://www.england.nhs.uk/aac/what-we-do/how-can-the-aac-help-me/ai-award/evaluation-for-artificial-intelligence/\)](https://www.england.nhs.uk/aac/what-we-do/how-can-the-aac-help-me/ai-award/evaluation-for-artificial-intelligence/) and shared with NHS England, NICE, and the UK National Screening Committee (where relevant). Based on the interim reports, guidance was provided on priorities for the final results. For example, an interim report on AI to support treatment decision-making in stroke care was not clear whether the observed increase in mechanical thrombectomy rates was due to implementation of the AI or reflected a nationwide increase. The Evaluation Advisory Group said this should be investigated in the final report.

A key lesson learned at this stage was:

1. Show trends in impact but note limitations and issues with generalising the results

The interim reports highlighted early impact trends to help national programme teams and NICE understand the potential impact of the AI technologies and to highlight gaps in understanding to be prioritised for the final report.

It was important to point out the limitations of the interim findings, highlighting that early positive trends might not be borne out over the longer term

Baseline analyses indicated significant variation in clinical practice and IT systems between sites. This underlined the importance of analysing differences in outcomes between sites and commenting on how generalisable results might be if an AI technology were to be scaled nationally.

Acknowledgements

The AI in Health and Care Award ran from 2020 to 2024 and was part of the NHS AI Lab, a Department of Health and Social Care initiative that was part of the Government Major Projects Portfolio. It was initially delivered and operated by the Accelerated Access Collaborative (AAC), in collaboration with the National Institute for Health Research (NIHR) before being handed to the NHS AI lab in 2023 to manage its close-out.

Work capturing lessons learned on real world AI evaluation from the AI Award was commissioned by NHS England and delivered by Unity Insights through a literature review, review of AI Award documents, interviews, and focus groups in early 2023. This NHS England report summarises and adds to those findings.

The authors would like to thank the members of the Accelerated Access Collaborative, NHS AI Lab, and the AI Award Evaluation Advisory Group who reviewed drafts of this report and the members of the AI Award evaluation teams who participated in interviews and focus groups.

Appendix A – case studies

Case study 1

Challenge

The evaluation looked at an AI tool to support the clinical assessment of CT scans and decisions about the appropriate treatment for stroke patients.

This technology and a competitor AI technology had already been implemented at some sites and there was resistance to switching technologies or to dealing with delays introduced by controlled trial randomisation.

There was a high degree of variation between the clinical models at hospital sites (for example, acting as ‘hubs’ or ‘spokes’ within the Integrated Stroke Delivery Network). There were also significant differences in the extent of access to, and baseline use of, key imaging modalities and treatment mechanisms (CT angiography, CT perfusion imaging, and mechanical thrombectomy).

Approach

The evaluation team designed a quasi-experimental (<https://www.gov.uk/guidance/quasi-experimental-study-comparative-studies>) pre-post intervention study to assess the effectiveness of the AI technology in reducing time to treatment decisions and in increasing treatment rates.

Given the variation between clinical practice at sites and potential differences in patient demographics, the evaluator planned a 6-month period of baseline profiling. They planned to map the clinical pathways and processes at each site against the NICE clinical pathway and to analyse patient demographics and outcomes.

The evaluator also planned to analyse outcomes at a selection of sites

without the technology that were matched to a selection of the 34 sites included in the project. This would help identify and account for trends outside the study that might affect the results (for example, a national increase in the use of mechanical thrombectomy).

A mixed methods approach was adopted, including a qualitative assessment of clinicians' experience with the technology (for example, which platform components the clinicians found useful to support their decision-making and why).

Case study 2

Challenge

This project involved an AI technology used to analyse chest images. The evaluation plan involved 4 work packages.

1. An accuracy workstream comparing the AI output for 200 scans with expert manual ratings of the same images.
2. Qualitative research with clinicians and patients to understand the acceptability of the technology.
3. A prospective, pragmatic, comparative effectiveness trial at 8 sites (4 would deploy the technology and 4 would not). Outcome pre- and post-deployment would also be compared.
4. The development of a cost-effectiveness model for the technology (fed by effectiveness data from work package 3).

During the implementation of the evaluation, there were delays in the recruitment of deployment sites for work package 3.

2 sites were recruited but it was noted that they were sending lower than expected numbers of images to the technology company. Feedback indicated that clinicians were not using the AI technology outputs to inform patient care.

Approach

The evaluation was redesigned to make work package 3 a retrospective study comparing AI outputs with real world clinical assessments (with

links made to patient outcomes). This would be combined with a survey of 60 clinicians to understand how they would use the AI technology in practice. As part of this, clinicians would be asked, hypothetically, how they would act if they disagreed with the AI outputs.

The outputs of this research would be combined with 'willingness-to-pay' data to inform the development of the cost-effectiveness model (work package 4).

The budget released by not delivering work package 3 as planned meant that work package 2 could be expanded to include a process evaluation to identify the barriers and enablers to implementation and use of the AI technology at sites.

Appendix B: evaluation questions and associated methods

This section shows what evaluation domains, methods of analysis and outcomes or metrics were associated with evaluation questions shared across projects.

Evaluation question	No. of evaluations	Evaluation domains	Analysis methods	Outcomes or metrics
---------------------	--------------------	--------------------	------------------	---------------------

Evaluation question	No. of evaluations	Evaluation domains	Analysis methods	Outcomes or metrics
How safe is the technology and what are the key risks or assurances?	7	Safety; accuracy; effectiveness; fit with sites; implementation; feasibility of scaling up; sustainability	Desk-based review of documentation Observational review of feedback sessions Interviews with key users Data mining to identify safety issues Comparator group analysis Systems analysis (for example, e-risk control, post deployment monitoring, hazard identification)	Compliance with safety standards Frequency of clinical deviation from AI recommendation Patient satisfaction Algorithmic error event causation Incidence of false positives and false negatives

Evaluation question	No. of evaluations	Evaluation domains	Analysis methods	Outcomes or metrics
How accurate is the technology compared to current care?	9	Safety; accuracy; effectiveness; implementation; feasibility of scaling up	Comparator group analysis of decisions AI validation using a machine learning expert Knowledge discovery and data mining Clinical validation Probabilistic sensitivity analysis	Throughput of patients Recall rate and precision Sensitivity and specificity Incidence of false positives and false negatives Frequency of clinical deviation Identification of erroneous assignments

Evaluation question	No. of evaluations	Evaluation domains	Analysis methods	Outcomes or metrics
Does implementing the technology make the pathway more efficient for staff (reduction in time) or for patients (reduction in time to treatment or diagnosis)	7	Safety; accuracy; effectiveness; value; feasibility of scaling up; fit with sites; sustainability	Time-motion study Comparator group analysis Surveys, interviews, or focus groups with clinicians and patients	Throughput or resource use (referrals, appointments , unnecessary or missed referrals or appointments) Clinician time spent Time to decision, referral or treatment Incidence of false positives and false negatives Uptake and use by clinicians device failures Patients' or clinicians' experience

Evaluation question	No. of evaluations	Evaluation domains	Analysis methods	Outcomes or metrics
Does implementing the technology impact patient outcomes?	7	Safety; accuracy; effectiveness; sustainability	Comparator group analysis Health inequalities analysis Qualitative interviews and focus groups Observational data surveys of GPs or administrators about use or changes in management	Throughput or resource use (for example, bed days, inappropriate referrals)time to decision, referral or treatment Device failures Clinical decisions or outcomes patient experience

Evaluation question	No. of evaluations	Evaluation domains	Analysis methods	Outcomes or metrics
Is there a valid economic case for the technology?	13	Value; Implementation; Feasibility of scaling up; Sustainability	Cost-benefit analysis Cost-effectiveness analysis Cost-utility analysis Budget impact analysis Cost-consequence analysis	Clinical or patient-reported outcomes Resource use Time to decision, referral or treatment Cost of resources (staff, bed days, technology)Health-related quality of life (HRQOL) improvement

Evaluation question	No. of evaluations	Evaluation domains	Analysis methods	Outcomes or metrics
Looking at the current pathway and the proposed pathway, is there variation between different demographics of patients in incidence, access or outcomes ?	2	Safety; accuracy; effectiveness	Comparator group analysis Health inequalities analysis Qualitative interviews with clinicians or patients	Time to decision, referral or treatment Resource use Patients' or clinicians' experience Throughput or resource use (referrals, appointments, unnecessary or missed referrals or appointments) Clinical outcomes or decisions Access Unwarranted variation in outcomes according to patient characteristics

Evaluation question	No. of evaluations	Evaluation domains	Analysis methods	Outcomes or metrics
How do patients or clinicians perceive and use the technology?	10	Safety; effectiveness; fit with sites; implementation; feasibility of scaling up; sustainability	Surveys, interviews or focus groups with clinicians and patients Comparison of recommended management outcomes and clinicians' decisions Ethnographic methods or observations	Patients' or clinicians' experience Trust in or acceptability of technology Aided or unaided patient management decision Perceived accuracy of results Patient demographics Compliance

Evaluation question	No. of evaluations	Evaluation domains	Analysis methods	Outcomes or metrics
What is the sustainability of the technology within the deployment site and in other deployment sites	3	Sustainability	Comparator group analysis Interviews and focus groups with patients and clinicians	Maintenance of CE Mark and ISO certifications Sensitivity, specificity, arbitration rate, recall rate, attendance rate Evolution of patient needs in relation to technology use
What should be considered when implementing the technology?	2	Implementati on; sustainability	Interviews and focus groups with patients and clinicians	Barriers and enablers to technology implementati on Training requirements Integration costs

Evaluation question	No. of evaluations	Evaluation domains	Analysis methods	Outcomes or metrics
What is the scale-up strategy?	1	Feasibility of scaling up	Interviews and focus groups with patients and clinicians engagement with the technology	Perceived impact of the technology implementation on staff, the organisation, and patients Patient demographics (to understand generalisability)
What is the real-world impact on clinical decision making of the technology?	2	Safety; effectiveness; fit with sites; sustainability	Interviews with clinicians Analysis of AI-recommendations versus clinical decisions	Aided or unaided patient management recommendations Clinicians' experience

Appendix C: national guidance and resources

Rapid changes in the sector meant teams leant heavily on national guidance and resources. Key sources of guidance and support are:

AI and Digital Regulations Service for health and social care (<https://www.digitalregulations.innovation.nhs.uk/>): Learn what regulations to follow and how to evaluate effectiveness, whether you're a 'developer' of AI and digital technology or an 'adopter' who will buy or use them in health and social care.

Provides guidance and case studies, produced by the Medicines and Healthcare products Regulatory Agency (MHRA), Health Research Authority (HRA), National Institute for Health and Care Excellence (NICE), and Care Quality Commission (CQC).

Office for Health Improvement and Disparities guides to evaluating digital health products (<https://www.gov.uk/government/collections/evaluating-digital-health-products>): The Office for Health Improvement and Disparities' guides are intended for anyone developing or evaluating a digital health product.

NHS Innovation Service resources (<https://innovation.nhs.uk/about-the-service/>): Registration with the NHS Innovation Service allows access to direct advice from partners including NICE, the National Institute for Health and Care Research (NIHR), and NHS England.

Appendix D: theoretical frameworks for the evaluation of AI technologies in health and care

This section provides a concise overview of the theoretical frameworks that are relevant to designing AI evaluations across the award's 4 phases:

Phase 1: AI design, training, and feasibility testing National Diabetes Experience Survey (<https://www.england.nhs.uk/diabetes/national-diabetes-experience-survey/>)

- TRIPOD+AI: updated guidance for reporting clinical prediction models that use regression or machine learning methods (<https://www.bmj.com/content/385/bmj-2023-078378>).
- The STARD-AI protocol: Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies (<https://bmjopen.bmj.com/content/11/6/e047709>).

Phase 2: Further AI development and clinical validation

- DECIDE-AI: Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence (<https://www.bmj.com/content/377/bmj-2022-070904>).

Phase 3: First prospective real-world deployment of the AI

- CONSORT-AI extension: Reporting guidelines for clinical trial reports for interventions involving artificial intelligence (<https://www.bmj.com/content/370/bmj.m3164>).
- SPIRIT-AI extension: Guidelines for clinical trial protocols for interventions involving artificial intelligence (<https://www.bmj.com/content/370/bmj.m3210>).

Phase 4: Multi-site deployment and real-world evaluation

- TEHA: Evaluation framework to guide implementation of AI systems into healthcare settings (<https://informatics.bmj.com/content/28/1/e100444>)
- NICE evidence standards framework (ESF) for digital health technologies (<https://www.nice.org.uk/about/what-we-do/our-programmes/evidence-standards-framework-for-digital-health-technologies>) and Real World Evidence framework (<https://www.nice.org.uk/corporate/ecd9/chapter/overview>)
- The Magenta Book: HM Treasury guidance on what to consider when designing an evaluation (<https://www.gov.uk/government/publications/the-magenta-book>)

This publication ('Planning and implementing real-world AI evaluations: lessons from the AI in Health and Care Award') is designed to complement the frameworks and provide practical support on the design and implementation of evaluations at phase 4.

Appendix E: evaluations covered by this review

Round	Company	AI technology	Technology use case	Evaluator
1	Aidence	Veye	Veye Lung Nodules (VLN) was intended to assist radiologists in their review of CT scans for the detection and classification of potential lung cancer nodules.	University of Edinburgh/ Hardian Health

Round	Company	AI technology	Technology use case	Evaluator
1	Brainomix Ltd	e-Stroke Suite	Enabling the sharing of CT scan images between clinicians and supported decision-making about the best treatments for stroke patients.	Oxford HIN
1	Deloitte	RITA: Referral Intelligence and Triage Automation	Triaging outpatient referrals from primary care into secondary care, reducing clinical administrative burden and helping clinicians spend more time with patients.	Unity Insights/ University of Surrey

Round	Company	AI technology	Technology use case	Evaluator
1	Healthy.io (UK) Ltd	Minuteful Kidney	Allowing people who were at higher risk of kidney disease to test and analyse their urine from home and see results using their mobile phone.	Midlands and Lancashire CSU
1	ICNH Ltd	DrDoctor	Supporting hospitals to predict who will not attend their appointment and intervening accordingly (for example, by phone call or reminder text message).	London South Bank University
1	iRhythm Technologies Ltd	Zio patch	A wearable patch with AI technology that aimed to detect irregular heart rhythms quickly and accurately.	King's Technology Evaluation Centre

Round	Company	AI technology	Technology use case	Evaluator
1	Kheiron Medical Technologies	Mia Mammography Intelligent Assessment	Analysing mammograms and supporting detection of breast cancer in screening services. It was designed to replace the second 'reader' of the mammograms, increasing confidence and reducing missed cancers.	King's Technology Evaluation Centre
1	Optos PLC	Optos AI	Using machine learning to analyse images of the retina and supporting screening and identification of eye disease in people with diabetes.	Midlands and Lancashire CSU

Round	Company	AI technology	Technology use case	Evaluator
2	Ultromics Ltd	EchoGo Pro	An AI algorithm that analysed stress echogram images (images of the heart when under stress) to support the diagnosis of heart disease.	London South Bank University
2	Skin Analytics Ltd	DERM	Using machine learning to analyse images of the skin and support clinicians in primary and secondary care to decide whether to refer people for cancer diagnostic testing.	Unity Insights/ University of Surrey

Round	Company	AI technology	Technology use case	Evaluator
2	eConsult Health Ltd	eHub	Using natural language processing and machine learning to analyse requests for GP appointments and supporting scheduling and decisions about the kinds of appointment required (for example, face-to-face or telephone).	Swansea University
2	University of Oxford	Paige prostate cancer detection tool	Analysing prostate images and supporting the identification and assessment of prostate cancer.	University of York

Round	Company	AI technology	Technology use case	Evaluator
2	Zebra Medical Vision	HealthVCF	Supporting detection of spine fractures in CT images of people aged over 50 years.	King's Technology Evaluation Centre

Publication reference: PRN01340

Date published: 16 October, 2024

Date last updated: 16 October, 2024

[▲Back to top](#)